# 数据期刊同行评议视角下科学数据质量评价指标识别*

■ 撤旭[1] 王健[1] 范智萱[1] 刘建平[1] 张贵兰[2] 徐波[3]
[1] 中国农业科学院农业信息研究所 北京 100081 [2] 中国科学技术信息研究所 北京 100038
[3] 国家科技基础条件平台中心 北京 100038

摘 要：[目的/意义]在数据期刊同行评议的视角下识别并构建科学数据质量评价指标，增强对科学数据质量评价的理解，为数据论文同行评议实践提供参考。[方法/过程]利用扎根理论的研究方法，选取 20 个数据期刊的数据同行评议指南作为质性研究的原始资料，并使用 NVivo 质性分析软件对资料进行开放式编码、关联式编码和选择性编码，通过理论饱和度检验对编码结果进行检验。[结果/结论]最终构建数据论文同行评议情境下的科学数据质量评价指标体系，识别出数据内在质量、数据表达质量、数据访问质量和数据效用质量 4 个主范畴和 14 个评价指标，并具体分析指标的含义及分类，帮助数据论文作者和评审者深入了解科学数据质量的内在结构。

关键词：数据期刊 科学数据质量 同行评议 评价指标 扎根理论
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.17.013



在学术界不断推进开放科学、科学数据共享以及数据密集型科学探究的背景下，科学数据共享的体量和学科领域覆盖度持续增加，制约科学数据共享和复用效益发挥的瓶颈因素逐渐转向科学数据的质量[1-2]。如何有效评价进而控制科学数据质量成为科学数据出版者、科学数据中心以及其他科学数据传播参与者共同关注的问题[3-4]。

在不同类型和形式的科学数据质量评价实践中，数据期刊开展的同行评议发展最为迅速，并且得到了学术界的普遍认同[4]。数据论文同行评议的被接受程度及其吸引的学术关注度展现了其良好的发展前景，成为其他类型科学数据质量评价可借鉴和参考的"最佳实践"。因此，研究数据论文同行评议实践，调研同行评议指南并从中提炼出评价内容和评价指标，既可以更深刻地了解其现状，也可以为科学数据中心等其他机构开展科学数据质量评价提供更有效的参考。

国内外学者已经围绕科学数据质量评价和数据论文同行评议开展了多项研究，涉及科学数据质量的构成、评价标准和评价方法[5-9]，以及数据论文的概念、性质、质量评价与控制[10-13]等多个主题。然而在文献分析时却发现，针对数据论文同行评议开展的研究主要使用调查、典型案例分析、综述和分析探讨等方法，这些研究未能完整地覆盖各类数据期刊，同时其研究结论缺乏必要的数据支撑。为此，本论文力图在全面调查现有数据期刊的基础上，以各期刊提供的同行评议指南作为分析对象，采用扎根理论方法识别科学数据质量评价指标，同时建立包括标准以及标准所属类别的评价框架，为数据期刊和科学数据中心等发展自身的评价标准体系提供参考。

## 1 文献综述

### 1.1 科学数据质量评价相关研究

R. Y. Wang 等是数据质量研究领域的主要代表性学者之一，他认为数据质量是指"适合数据消费者使用的数据(Fit-for-use)"，可以通过准确性和完整性等维度予以表征和度量。他从用户实验中得到数据质量的概念框架，框架中包含内在数据质量、可访问的数据质量、上下文数据质量、表达性数据质量 4 个质量方面

* 本文系中国农业科学院农业信息研究所基本科研业务费重点项目"科学数据出版能力建设研究"(项目编号:JBYW-AII-2020-35)和国家科技基础条件平台专项课题"科学数据质量评价研究"(项目编号:2018DDJ1ZZ16)研究成果之一。

作者简介：撤旭(ORCID:0000-0001-9493-4960)，硕士研究生，E-mail:saxu517@163.com；王健(ORCID:0000-0003-4958-7669)，研究员，博士；范智萱(ORCID:0000-0002-1906-8174)，硕士研究生；刘建平(ORCID:0000-0002-1817-5373)，博士研究生；张贵兰(ORCID:0000-0002-9153-3579)，助理研究员，博士；徐波(ORCID:0000-0001-8262-579X)，助理研究员，博士。

收稿日期:2020-03-13 修回日期:2020-04-18 本文起止页码:123-130 本文责任编辑:易飞

和 15 个数据质量维度[14]。该框架被 P. Katerattanakul、B. Klein 等许多学者验证和使用[15-16]，为后续科学数据质量评价的研究奠定了基础。国内外学者从数据生产、使用和管理等多个方面对科学数据质量控制和评价展开了相应的研究。L. Waaijers 等从数据的生命周期方面调查了应该对研究性数据进行的质量控制操作，发现在数据生产阶段，设备的准确性和方法问题非常重要；在数据管理阶段，要确保数据的持久性和可访问性，其中良好的元数据是一个基本要求；在数据重用阶段，要关注数据的实际内容质量，即数据集的学术价值[5]。V. Lush 等对地理空间数据的用户和专家进行了一系列半结构化访谈，以确定地理空间数据集的关键质量指标，研究发现拥有完整元数据记录、良好声誉的数据集更可能被用户视为“高质量”的数据，并且在评价数据质量时，用户主要依赖于元数据内容质量、元数据的可视化、评审者意见、数据提供者的信誉、引文信息和数据创建者提供的软知识（建议、评论等）[6]。M. Gamble 等发现科研人员使用数据时关注数据的质量（与规范和标准相比是否良好）、信任（数据来源可能不错）和效用（数据适应当前需要）三个方面[7]。崔旭等从科学数据管理的角度对数据选择与评价的关键要素进行了研究，总结出安全性、精确性、相关性、可获得性和可用性 5 个数据质量评价标准[8]。屈文建等通过分析高校科研数据质量存在的问题，构建了数据质量控制架构体系，指出要从准确性、完整性、合时性和一致性来评估数据质量[9]。

综合分析以上研究发现，学者们普遍认为可以通过多个质量指标来对科学数据质量进行评价，以便对数据的各个方面进行质量把控。但是从不同研究角度出发，学者们对科学数据质量内涵的理解以及提出的评价指标都存在差异，有必要针对具体的评价情境和评价主体对科学数据质量评价指标进行更为深入的探索。

### 1.2 科学数据质量同行评议指标相关研究

同行评议是一个或多个专业知识和经验丰富的学科领域专家共同对科学数据质量进行评价的过程。从同行评议中可以获得数据集的优点、数据集的问题和其他信息反馈[5]。科研人员普遍认为经过严格同行评议的科学数据具有一定的可信度和声誉。因此，在数据同行评议成为必然趋势的背景下，越来越多的学者开始关注如何对科学数据质量进行同行评议以及应该采用哪些指标进行评议。

2014 年，荷兰莱顿“联合共建数据公平港口”学术研讨会上，代表学术界、产业界、资助机构和学术出版商的不同利益相关者集合在一起，提出并共同认可“可发现（findable）、可访问（accessible）、可互操作（interoperable）和可重用（reusable）”的 FAIR 原则[17]，为数据管理和数据发布提供了广泛的准则，为数据出版中的数据评审标准的建立提供了主要依据。根据数据出版的特点，国内外学者对科学数据质量同行评议指标展开了研究。J. E. Kratz 等探讨了科研人员对数据出版和同行评审的期望，发现研究者们希望同行评审关注的 7 个方面，分别是方法是否恰当、可重用性、技术质量、数据可信度、元数据是否标准、新颖性和影响力[18]。T. A. Carpenter 通过调研得到数据论文的 4 类评审标准，分别是编辑审查标准、元数据质量标准、数据质量标准、方法审查标准[4]。国内学者中，刘传玺选取 10 种代表性数据期刊进行调研，通过分析其数据论文的同行评审指南，总结得到 5 个方面的评审标准，分别为论文质量控制、数据方法的质量、文章和数据的一致性、数据的可用性、数据的效用和价值[10]。涂志芳指出科学数据中存在科学性、技术性和监护性 3 类评审，其中科学性评审是对数据内容的科学性特征进行评审，表现为完整性、准确性、真实性、有用性、可靠性等特征，通常由同行评审专家进行评审[11]。孔丽华等基于 FAIR 数据共享原则对数据出版中数据质量评价指标进行设计，得到可获取、可评估、可理解和可重用等 4 个一级指标和 13 个二级指标[12]。李晓蕾等根据地质科学数据的特点指出，应该在地质科学数据出版过程中对完整性、可用性、专业性、保密内容、敏感内容、公开发表内容 6 个方面进行质量审查[13]。

通过文献调研发现，科学数据质量同行评议指标的研究处于探索性阶段。国内外相关学者主要以总结归纳的方式提出科学数据质量评价指标，并且指标之间差异较大，还未形成统一的评价标准和框架，尤其是现有研究中评价指标是否全面、指标内涵是否明确等问题仍有待讨论。因此，本研究以数据期刊已经发布的数据评审指南为基础资料，通过扎根理论对其中的质量评价标准进行定性分析，以期发现并构建科学数据质量同行评议指标体系，为评审专家和其他数据用户判断科学数据质量提供参考。

## 2 研究设计

### 2.1 研究方法

本文采用扎根理论方法构建科学数据质量同行评议指标体系。扎根理论是由社会学家 B. G. Glaser

等[19]所提出的定性研究方法,为形成和处理丰富的定性材料提供了系统的程序。利用扎根理论方法可以从收集的材料中逐步发展出更抽象的概念范畴,并确定其中的模式关系。关于科学数据质量概念的界定还没有得到学术界的共识,其评价指标、评价内容以及两者的对应关系在以往的研究中也存在矛盾,有必要调研数据期刊同行评议指南的内容现状并从这些现象中总结出科学数据质量的内涵和评价指标,丰富科学数据质量的理论研究。因此,本研究适合采用扎根理论来对科学数据质量同行评议指标和范畴进行质的分析和确定。

## 2.2 数据收集

研究以数据期刊的同行评议指南或类似政策文本作为分析对象。数据期刊界定为曾发表过数据论文的期刊,包括只发表数据论文的纯数据期刊和同时发表其他类型论文的混合型数据期刊。研究首先通过两个途径确定数据期刊列表,然后根据列表逐一获取其同行评议指南或类似政策文本,同时根据文本获取情况再次精炼期刊列表。在第一个数据期刊获取途径中,首先通过 Web of Science 检索确定数据论文,然后提取不同论文所在的期刊。检索首先为主题过滤,其检索条件为:"主题:(in) OR 主题:(on) OR 主题:(by) OR 主题:(at) OR 主题:(about) OR 主题:(under) OR 主题:(of) OR 主题:(the)",其次按照文献类型选择"Data Paper"进行过滤。如此得到 2 398 篇数据论文,分别来自 95 个不同的期刊。在第二个途径中,主要参考 L. Candela 等在 2015 年统计的一个包含 15 家出版机构出版的 116 种数据期刊的清单[20],以及刘凤红等在 2019 年对 L. Candela 清单进行更新、扩展后所形成的包括 26 家出版机构和 168 种数据期刊在内的清单[21]。

基于以上清单,对数据期刊或其所属出版机构官方网站上的评审指南或类似政策文档进行核验,一方面根据核验情况从清单中排除重复(因若干期刊属于同一出版集团而共享相同的策略)或缺乏相关文档的期刊,一方面下载或复制评审指南等相关文本。所有英文文本均翻译为中文,并通过作者之间的交叉检验保证翻译质量。通过上述步骤,研究最终确定了 20 份来自不同期刊的同行评审指南,数据期刊的详细信息见表 1,其中 J1-J11 为混合型数据期刊,J12-J20 为纯数据期刊。

表 1　数据期刊详细信息

| 编号 | 数据期刊名称 | 所属出版商或出版集团 | 所属学科领域 |
|---|---|---|---|
| J1 | *F1000 Research* | F1000 Research | 综合性 |
| J2 | *Ecology* | Ecological Society of America | 环境与生态学 |
| J3 | *GigaScience* | Oxford University | 生物学、医学 |
| J4 | *Biodiversity Science* | 中国科学院生物多样性委员会等 | 生物学 |
| J5 | *BMC Research Notes* | Biomed Central | 综合性 |
| J6 | *Ecological Research* | Wiley | 生物学 |
| J7 | *Earthquake Spectra* | Earthquake Engineering Research Institute | 工程技术、地质 |
| J8 | *Advances in Atmospheric Sciences* | Science Press | 大气和物理海洋学 |
| J9 | *Genetics* | Frontiers | 遗传学 |
| J10 | *Data Science Journal* | CODATA | 综合性 |
| J11 | *PlosONE* | Plos | 综合性 |
| J12 | *Scientific Data* | Springer-Nature | 综合性 |
| J13 | *Data in Brief* | Elsevier | 综合性 |
| J14 | *Earth System Science Data* | Copernicus | 地球科学、气象与大气科学 |
| J15 | *Geoscience Data Journal* | Wiley | 地球科学、气象与大气科学 |
| J16 | *Biodiversity Data Journal* | Pensoft | 生物学 |
| J17 | *Open Health Data* | Ubiquity | 医学 |
| J18 | *Data* | MDPI | 综合性 |
| J19 | 中国科学数据 | 中国科学院计算机网络信息中心 | 综合性 |
| J20 | 全球变化数据学报 | 中国科学院地理科学与资源研究所 | 地球科学、气象与大气科学 |

## 2.3 研究过程

基于扎根理论的研究思路,利用 Nvivo11 软件对原始资料中的语句进行开放式编码、关联式编码和选择性编码。本研究在 20 份同行评审指南中选择 17 份作为分析对象,预留 3 份数据期刊(J9、J10 和 J11)的同行评审指南作为校验样本。为减少编码的主观性,本研究参考

R. Y. Wang 等[14]提出的数据质量概念框架，建立了科学数据质量的概念框架，包括数据内在质量、数据表示质量、数据访问质量和数据效用质量 4 个方面，如表 2 所示。利用该框架对从原始资料中提炼出的科学数据质量指标进行分类，同时利用扎根理论在实际编码过程中对框架中的要素内涵进行适当扩充和修改。

表 2 科学数据质量概念框架

| 分类 | 含义 |
|---|---|
| 数据内在质量 | 数据集固有的质量维度，包括准确性、真实性、有效性等内容 |
| 数据表示质量 | 数据集描述信息的清晰、准确、完整、一致和易理解程度 |
| 数据访问质量 | 数据集及其描述信息的可发现和可获得的程度 |
| 数据效用质量 | 数据集在特定或通用情境中的作用价值 |

#### 2.3.1 开放式编码

原始数据的开放式编码可实现逐层的概念化和范畴化，概念是编码的最小意义单元，范畴是概念抽象后所表现的观点或主题[22]。在开放式编码中，先将收集到的所有同行评审标准划分为 181 条原始语句（参考点），然后对原始语句进行精炼并且将意思相同的语句提炼为同一个概念，例如将“是否使用适当的方法来收集和处理数据？（J7-3）”和“数据产生方法是否适宜？（J20-5）”都提炼为“方法适当（a15）”，共计得到 87 个初始概念。之后通过不断对比分析，将具有共性的初始概念进行合并从中抽象出 14 个范畴，部分编码过程如表 3 所示。开放式编码由 2 位编码员共同完成，利用 O. R. Holsti[23]可信度公式对编码结果进行一致性检验，结果显示一致性程度为 82%。针对检验中不一致的情况，由课题组共同商议决定，最终编码结果如表 4 所示。

表 3 部分开放式编码过程

| 原始语句 | 初始概念 | 范畴 |
|---|---|---|
| J16-6 数据资源是否涵盖了足够大的区域、时间段，值得出版？ | a1 空间和时间范围的满足 | A1 完整性 |
| J18-7 是否适当描述了可能的错误源？ | a9 适当讨论错误源 | A2 准确性 |
| J20-5 数据产生方法是否适宜？ | a15 方法适当 | A3 可靠性 |
| J8-3 是否对数据集进行了定期更新 | a27 数据定时更新 | A4 及时性 |
| J2-3 表达简明扼要，容易理解吗？ | a28 表达简明；a29 表达容易理解 | A5 易理解性 |
| J7-5 数据和手稿是否基本一致？ | a32 论文和数据的一致性 | A6 一致性 |
| J13-3 数据格式是标准的吗？ | a40 数据格式符合规范 | A7 规范性 |
| J3-4 是否为应提交学术社区认可的公共储存库的数据提供了链接？ | a50 提供数据库链接 | A8 可访问性 |
| J8-8 数据集易于下载吗？ | a57 数据容易下载 | A9 可获取性 |
| J12-8 这些数据文件是否被存放在了最合适的数据知识库中？ | a60 合适的数据存储库 | A10 可存储性 |
| J15-4 这些数据对地球科学做出了重要而独特的贡献吗？ | a63 对学科领域有贡献 | A11 增值性 |
| J18-10 数据可重复使用吗？ | a77 数据可重用 | A12 可重用性 |
| J14-1 数据和方法是新的吗？ | a81 方法新颖 | A13 新颖性 |
| J20-3 是否前人已有相同的数据发表？ | a85 没有重复发表数据 | A14 唯一性 |

#### 2.3.2 关联式编码

关联式编码实现范畴性质的挖掘和范畴与主范畴间关联关系的发现，同时对范畴进行重新归类和融合。根据科学数据质量概念框架，将开放式编码形成的 14 个范畴归类到数据内在质量、数据表达质量、数据访问质量和数据效用质量 4 个主范畴中（见表 5）。

#### 2.3.3 选择性编码

选择性编码指在所有已发现的概念类属中系统分析并总结形成一个“核心类属”，以此连贯整个编码过程，从而构建理论模型。分析编码过程可以发现，4 个主范畴分别对应同行评议对科学数据质量不同方面的要求，因此提炼出“数据期刊同行评议视角中的科学数据质量”这一核心范畴。

#### 2.3.4 理论饱和度验证和指标体系构建

依照上述编码过程对预留用作理论饱和度检验的 3 份数据评审指南进行再次编码，发现编码结果都能列入已形成的编码概念之中，未发现新的概念、范畴和典型关系，证明编码实现了理论饱和且核心范畴有效。最终，本文立足于数据期刊同行评议的视角，通过编码结果构建了科学数据质量评价指标体系，见图 1。

## 3 科学数据质量评价指标分析

### 3.1 数据内在质量

从编码结果来看，数据内在质量不仅指数据值与实际值或真实值一致的程度，也指数据的完整程度和更新程度，可以通过准确性、可靠性、完整性和及时性 4 个指标进行评价。可靠性是数据内在质量中出现频

**表4　开放式编码结果**

| 范畴(参考点) | 初始概念 |
|---|---|
| A1 完整性(25) | a1 空间和时间范围的满足;a2 数据结构完整;a3 数据缺失;a4 数据完整;a5 数据文件的完整性;a6 元数据完整;a7 足够的数据量;a8 足够的数据深度 |
| A2 准确性(9) | a9 适当讨论错误源;a10 数据准确;a11 数据异常值被良好记录;a12 说明数据误差限制 |
| A3 可靠性(32) | a13 方法和仪器先进;a14 方法科学;a15 方法适当;a16 方法严格;a17 良好的数据质量控制;a18 适当引用相关数据集和文章;a19 数据技术合理;a20 数据来源明确;a21 数据没有作假;a22 数据样本具有代表性;a23 数据样本体量适当;a24 数据有逻辑;a25 数据真实可靠;a26 正确的研究设计 |
| A4 及时性(2) | a27 数据定时更新 |
| A5 易理解性(4) | a28 表达简明;a29 表达容易理解;a30 恰当的表现形式 |
| A6 一致性(15) | a31 标题、摘要和关键词准确描述数据;a32 论文和数据的一致性;a33 数据产生方法和结果一致;a34 数据一致性;a35 数据与元数据一致;a36 正确描述数据 |
| A7 规范性(23) | a37 符合数据标准;a38 符合提交标准;a39 符合制度/公约/条款;a40 数据格式符合规范;a41 数据类型符合要求;a42 数据组织合理;a43 说明利益冲突和道德问题;a44 缩写和符号正确定义;a45 限制敏感数据的使用;a46 元数据的充分描述;a47 元数据符合标准;a48 元数据说明数据所有权;a49 元数据组织合理 |
| A8 可访问性(18) | a50 建立数据库链接;a51 数据公开提供;a52 数据可访问;a53 数据能随时提供;a54 提供唯一标识符 |
| A9 可获取性(7) | a55 适当的版权许可说明;a56 数据符合开放共享协议;a57 数据容易下载;a58 数据完全开放共享或协议共享;a59 说明如何获取数据和分析工具 |
| A10 可存储性(4) | a60 合适的数据存储库;a61 数据可永久保存 |
| A11 增值性(15) | a62 充分解释数据价值;a63 对学科领域有贡献;a64 具有发表意义;a65 具有科学意义;a66 数据创建理由和意义清晰;a67 数据能支持研究结论;a68 数据有使用价值 |
| A12 可重用性(35) | a69 方法便于重用;a70 方法描述详细;a71 分析工具的可用性;a72 实验可重复;a73 数据格式可重用;a74 数据和软件可使用;a75 数据可操作;a76 数据可用于其他实验或验证;a77 数据可重用;a78 数据描述详细充分;a79 提供合适的软件和服务;a80 提供数据重用建议 |
| A13 新颖性(6) | a81 方法新颖;a82 数据加工处理和质量控制过程创新;a83 数据具有新颖性;a84 数据来源创新 |
| A14 唯一性(3) | a85 没有重复发表数据;a86 没有重复实验或观察;a87 数据具有独特性 |

**表5　关联式编码结果**

| 主范畴 | 范畴 | 范畴内涵 |
|---|---|---|
| B1 数据内在质量 | A1 完整性 | 数据有足够的数量、广度和深度,提交的数据实体、元数据、数据文件的完整程度 |
| | A2 准确性 | 数据正确、无误的程度 |
| | A3 可靠性 | 数据的产生处理过程被接受或被认为是真实、可靠和可信的程度 |
| | A4 及时性 | 数据更新的程度 |
| B2 数据表达质量 | A5 易理解性 | 数据、数据表现形式及数据描述简明、清晰无歧义且易于理解的程度 |
| | A6 一致性 | 数据描述、元数据和数据实体的一致对应程度 |
| | A7 规范性 | (元)数据符合现行标准、公约、条例或规则的程度 |
| B3 数据访问质量 | A8 可访问性 | 数据能够通过唯一标识符和数据库链接进行快速检索的程度 |
| | A9 可获取性 | 数据易于下载、获取和查看的程度 |
| | A10 可存储性 | 使用恰当的数据库和数据长期保存的程度 |
| B4 数据效用质量 | A11 增值性 | 数据的有益程度和利用数据带来的好处 |
| | A12 可重用性 | 数据作者提供数据集的全部信息以便他人重复使用的程度 |
| | A13 新颖性 | 数据来源、产生方式、方法创新的程度 |
| | A14 唯一性 | 数据与已发表数据的重复程度 |

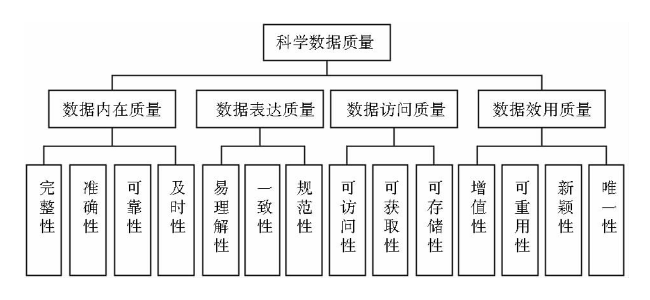



**图1　科学数据质量评价指标体系**

率最高的评价指标,涉及同行评审指南中的32个原始语句。数据可靠性评价体现的是评审者对数据的生产、处理和分析过程的综合判断,包括数据来源可靠、数据方法可靠、数据技术可靠等多个方面。例如原始语句中"数据是原始的,还是来源明确的?(J18-1)""有无发现或怀疑有数据作假情况(J20-11)"等是对数据来源是否可靠的判断;"数据收集方法是否具有较高的科学水平?(J8-2)"是评价数据方法的可靠程度;"数据集在技术上是否合理?(J18-5)"是评价数据技术方面的可靠性。数据作者对数据质量控制的描述和

说明也是评审者评价可靠性的判断依据,这可能是因为即使方法和工具非常先进,数据也总是不可避免地产生偏差或错误,需要作者采取一定的质量控制措施来保证数据的可靠性。完整性和可靠性有一定的关系,J. C. Wallis 等指出数据完整性与可靠性有关,如果数据的生产处理程序可靠,则数据可以在此过程中保持完整[24]。完整性也是数据内在质量中出现频率较高的评价指标,涉及同行评审指南中的 25 个原始语句。完整性的内涵除了屈文建[9]、R. Y. Wang 等[14]学者提出的数据广度、深度和结构完整程度以外,还包括元数据和数据文件的完整性,例如原始语句提到的"是否能够确认作者所存储的数据文档是完整的并与数据描述符中的描述相匹配?(J12-7)""元数据是否完整并足以促进数据的解释(J2-4)"。准确性和及时性都是出现频率非常低的评价指标。准确性涉及同行评审指南中的 6 个原始语句,可以从错误值、异常值和数据误差等方面进行判断,例如原始语句提到的"是否适当描述了可能的错误源?(J18-7)"。数据准确、无误虽然是数据质量的核心要素,但是评审者可能很难对其进行验证和判断,只能更多地关注可靠性和完整性等指标。数据的不断更新使得数据能够及时反映现实世界,因此,及时性是间接反映数据真实可靠的评价指标。以上 4 个评价指标的分析体现了数据内在质量是科学数据质量的关键和根本,能够反映数据收集、处理等各个阶段的质量情况。



### 3.2 数据表达质量

数据表达质量指数据以规范、一致和可理解的方式呈现的程度,可以从规范性、一致性和易理解性三个指标进行考量。一致性和可理解性在孔丽华[12]、R. Y. Wang[14]等学者的研究中都有提及,而规范性较少被发现。本研究中通过编码得到,规范性是数据表达质量中最受关注的评价指标,涉及到 23 个原始语句。M. Wendelbo 也指出如果数据被正确地标记和呈现,用户将更有可能理解和接受数据[25]。规范性的含义是指(元)数据符合现行标准、公约、条例或规则的程度。由于不同学科领域中数据具有多元化特征,学术界制定了不同专业领域的数据标准和元数据标准,以方便科学数据的管理和共享,并且可以减少评审者评价科学数据质量的主观性。从原始语句中"数据格式是标准的吗?(J13-3)""数学公式、符号、缩写和单位是否正确定义和使用?(J14-15)""元数据是否准确地描述了数据并遵守相关的学科或国际标准?(J18-3)""元数据在逻辑上是否有组织,它们是否符合元数据内容标准?(J2-1)"可以看出,评审者可以通过数据格式、缩写、符号、元数据格式、元数据结构等方面是否满足标准来判断规范性。一致性来源于编码资料中的 15 个原始语句。一致性不仅包括数据格式的一致性,也包括数据和元数据一致、数据与数据描述一致。易理解性受关注度较低,仅涉及 4 个原始语句,指数据表现形式及数据描述简明、清晰、无歧义且易于理解的程度。科学数据是反映事实并作为研究证据的特殊的传播对象,规范性保障科学数据的有序传播,一致性和易理解性帮助用户最大程度地理解和接收数据传递的信息。因此,良好的数据表达质量是科学数据共享的前提,是评审者从使用角度对数据质量的考量。

### 3.3 数据访问质量

数据访问质量不仅指数据方便用户访问、获取的程度,也包含数据实体的存储情况,可以通过可访问性、可获取性和可存储性三个指标来进行评价。可访问性是数据访问质量中最受关注的指标,涉及 18 个原始语句。从编码的原始语句"是否可以通过给定的标识符访问数据集(J14-5)""是否为应提交社区认可的公共储存库的数据提供了链接?(J3-4)"可以发现,数据库链接和标识符是评价可访问性的主要依据。J. E. Kratz 等指出研究人员希望出版的数据能够通过数据库或存储库被访问,数据期刊往往也建议和鼓励数据提交者能够公开他的数据[18]。C. C. Austin 等指出如果数据被分配 DOI 等持久性标识符,则用户可以快速方便地获取数据及其与其他出版物之间的联系[26]。确保数据的长期可访问性还涉及数据实体的存储,特别对于部分未建立独立数据存储机制的数据期刊,往往要求数据提交者选择适当的仓储以保证数据的持久存储和访问,例如原始语句中提到的"提交数据的存储库是否适合数据的性质?(J16-10)"等评价标准。可访问性仅指数据能够通过链接或标识符被检索和访问到,可获取性是指在数据可访问的前提下,数据还可以被查看、获取和下载[12]。J. E. Kratz 等在 2015 年的调查中指出,作为数据使用者的研究人员最关注出版数据的可获取性[18],在编码结果中也发现可获取性受到数据期刊同行评议的关注。除了数据是否容易下载,版权问题也是获取数据时需要考虑的因素,例如原始语句提到的"版权许可是否被描述(首选开放许可,但如果有令人信服的原因,作者可以使用其他许可)?(J18-4)"。大部分数据期刊将数据的访问质量作为重要的评审对象,反映了学术界对数据可获取和可访问的密切关注。

### 3.4 数据效用质量

数据效用质量指数据的可用程度和学术价值,可以通过可重用性、增值性、唯一性和新颖性 4 个指标进行评价。在编码得到的 14 个评价指标中,可重用性是出现频率最高的指标,符合数据期刊致力于科学数据共享和重用的理念。从编码的原始语句“是否描述了足够详细的数据收集方法,以允许另一位研究人员重现结果?(J18-2)”中发现,评价数据可重用的关键在于数据描述详细、全面的程度。评审者可以从分析工具的可用性、实验可重复、数据格式可重用、数据和软件可使用、数据可操作、数据可用于其他实验或验证等多个方面评价可重用性。增值性也受到同行评议的较高关注,涉及 15 个原始语句。增值性的含义是数据的有益程度和利用数据带来的好处[14],意味着对数据学术价值的判断,例如编码的原始语句中“这些数据对地球科学做出了重要而独特的贡献吗?(J15-4)”。评价增值性时,评审者以主观感受为主,不仅要考虑数据是否满足现行科研活动需要,还要从长远角度出发,准确预测数据是否可能满足未来的科学研究需要。然而对评审者而言,预测数据在未来如何被使用可能有一定的困难,可以借助作者对数据价值、数据创建理由和意义的描述来判断增值性。科学数据往往来源于科研人员的项目/课题研究,并且在作者发表研究性论文的时候被一并发表,因此,评审者也可以通过关注数据对研究结论的支持程度来评价其增值性。孔丽华[12]、李晓蕾[13]等提出的评价指标中都未包含数据唯一性和新颖性,编码结果也表明,唯一性和新颖性的受关注度较低。从原始语句“是否前人已有相同的数据发表?(J20-3)”“数据是否具有独特性(J20-2)”等中总结出唯一性的含义是指数据与已发表数据的重复程度,能够体现数据的发表价值。新颖性虽然不是数据必须具备的性质,但是可以体现数据的学术价值。如果数据来源新颖或数据方法、处理过程具有创新性,则评审者会容易判断数据是否有较高的科学意义。科学数据的本质是服务于科学研究的数据,在数据用户对数据重复利用的期望下,同行评议关注数据的效用质量,进而保障数据价值和作用的发挥。

## 4 结论与展望

本文以数据论文同行评议指南或类似政策文本为对象,通过扎根理论方法分析、提取科学数据的质量评价指标,以期从一个新的角度探索科学数据质量评价。通过研究,得到了数据内在质量、数据表达质量、数据访问质量和数据效用质量 4 个一级指标和完整性、准确性等 14 个二级指标,指出了部分指标的使用频次,构建形成了科学数据质量评价指标体系。最后,论文进一步讨论和分析了科学数据感知质量的内在结构以及质量指标和质量判断依据之间的关系。

在科学数据共享总量持续扩大的情况下,科学数据质量逐渐成为了有效共享的新短板,如何科学准确地评价科学数据质量是国内蓬勃发展的各类数据期刊和数据中心的迫切需求。在这一背景下,本文以实证的形式对国内外数据期刊的同行评议实践进行总结与提炼,一方面为相关各方展现科学数据质量评价现状,另一方面也为数据期刊和数据中心制定其数据质量评价指南和评价标准提供有益的参考。

论文在样本数量上存在一定的不足。两个因素导致了这一局限:一方面,数据论文仍然是学术传播领域的新生事物并且处于发展过程中,其总体数量相对有限;另一方面,部分期刊并没有公开其评审指南、部分期刊指南内容描述不充分以及存在评审指南共用(例如隶属同一出版集团的多个期刊使用一样的评审指南)等情况导致了大量期刊无法成为有效样本。论文通过理论饱和度校验部分地表明了当前样本数量并未对研究结论产生影响,但更大规模且学科覆盖范围更全面的样本将有助于进一步验证评价指标体系的通用性,这也是未来研究的重点和方向。

**参考文献:**

[1] GREGORY K, GROTH P, COUSIJN H, et al. Searching data: a review of observational data retrieval practices in selected disciplines[J]. Journal of the Association for Information Science and Technology, 2019, 70(5): 419 - 432.
[2] PRICE G. Figshare releases "the state of open data 2019" report [EB/OL]. [2020 - 04 - 14]. https://www.infodocket.com/2019/10/23/figshare-releases-the-state-of-open-data-2019/.
[3] 王丹丹. 科学数据出版过程中的数据质量控制[J]. 图书情报工作,2015,59(23):124 - 129.
[4] CARPENTER T A. What constitutes peer review of data: a survey of published peer review guidelines[DB/OL]. [2020 - 06 - 07]. https://arxiv.org/abs/1704.02236.
[5] WAAIJERS L, VANDERGRAAF M. Quality of research data, an operational approach[DB/OL]. [2020 - 03 - 11]. http://www.dlib.org/dlib/january11/waaijers/01waaijers.html.
[6] LUSH V, BASTIN L, LUMSDEN J. Geospatial data quality indicators[C]// Proceeding of the 10th international symposium on spatial accuracy assessment in natural resources and environmental sciences. Washington: International Spatial Accuracy Research Association, 2012:121 - 126.
[7] GAMBLE M, GOBLE C. Quality, trust, and utility of scientific data on the Web: towards a joint model[C]//Proceedings of the 3rd inter-

national Web science conference. New York: ACM, 2011: 1 – 8.
[8] 崔旭,韩子鹤,王铮,等. 科学数据管理中的数据选择与评价关键要素研究[J]. 情报理论与实践,2018,41(3):78 – 82,100.
[9] 屈文建,唐晶,陈旦芝. 高校科研数据质量控制架构与机制研究[J]. 情报理论与实践,2018,41(11):45 – 50.
[10] 刘传玺. 数据论文概念辨析及其同行评审研究[J]. 图书馆杂志, 2016,35(9):76 – 80.
[11] 涂志芳. 科学数据出版的基础问题综述与关键问题识别[J]. 图书馆, 2018, (6):86 – 92,100.
[12] 孔丽华,刁妍,郎杨琴,等. 数据期刊中科学数据的同行评议方法研究[J]. 编辑学报,2019,31(3):262 – 266.
[13] 李晓蕾,齐钒宇,孟洁,等. 地质科学数据出版的质量控制及公开化审查研究[J]. 中国矿业,2019,28(6):65 – 68.
[14] WANG R Y, STRONG D M. Beyond accuracy: what data quality means to data consumers[J]. Journal of management information systems, 1996, 12(4):5 – 33.
[15] KATERATTANAKUL P, SIAU K. Measuring information quality of Web sites: development of an instrument[C]//Proceedings of the 20th international conference on information systems. Atlanta: Association for Information Systems, 1999: 279 – 285.
[16] KLEIN B. When do users detect information quality problems on the World Wide Web? [C]//Proceedings of the 8th Americas conference on information systems. Atlanta: AIS Electronic Library, 2002: 1101 – 1103.
[17] WILKINSON M D, DUMONTIER M, AALBERSBERG I J J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific data, 2016, 3:160018.
[18] KRATZ J E, STRASSER C. Researcher perspectives on publication and peer review of data[J]. Plos one, 2015, 10(2): e0117619.
[19] GLASER B G, STRAUSS A L. The discovery of grounded theory: strategies for qualitative research[M]. Chicago: Aldine, 1967.
[20] CANDELA L, CASTELLI D, MANGHI P, et al. Data journals: a survey[J]. Journal of the Association for Information Science and Technology, 2015, 66(9): 1747 – 1762.
[21] 刘凤红,彭琳. 国际数据期刊的发展现状调查与分析[J]. 中国科技期刊研究,2019,30(11):1129 – 1134.
[22] 陈欣,叶凤云,汪传雷. 基于扎根理论的社会科学数据共享驱动因素研究[J]. 情报理论与实践,2016,39(12):91 – 98.
[23] HOLSTI O R. Content analysis for the social sciences and humanities [J]. New Jersey: Addison-Wesley, 1969, 14(11):137 – 141.
[24] WALLIS J C, BORGMAN C L, MAYERNIK M S, et al. Know thy sensor: trust, data quality, and data integrity in scientific digital libraries[C]//Proceedings of the 11th European conference on research and advanced technology for digital libraries. Berlin:Springer-Verlag, 2007: 380 – 391.
[25] WENDELBO M. Perspectives on peer review of data: framing standards and questions[J]. College & research libraries, 2017, 78(3): 262 – 266.
[26] AUSTIN C C, BLOOM T, DALLMEIER-TIESSEN S, et al. Key components of data publishing: using current best practices to develop a reference model for data publishing[J]. International journal on digital libraries, 2017, 18(2): 77 – 92.

**作者贡献说明:**
撒旭:论文数据采集、扎根理论编码分析和论文撰写;
王健:提出研究框架,指导论文修改;
范智萱:论文数据采集与编码校验;
刘建平:论文数据采集与编码校验;
张贵兰:提供论文修改建议与编码校验;
徐波:指导论文修改。



## Identification of Scientific Data Quality Evaluation Indicators from the Perspective of Data Journals Peer Review

Sa Xu[1] Wang Jian[1] Fan Zhixuan[1] Liu Jianpin[1] Zhang Guilan[2] Xu Bo[3]
[1] Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081
[2] Institute of Scientific and Technical Information of China, Beijing 100038
[3] National Science & Technology Infrastructure Center, Beijing 100038

Abstract: **[Purpose/significance]** From the perspective of peer review on data journals, this paper identifies and puts forward scientific data quality evaluation indicators to improve the understanding of scientific data quality evaluation and provide a reference for the practice of peer review of data papers. **[Method/process]** Data review guidelines for 20 data journals were selected as source material for qualitative research. This paper used grounded theory and qualitative analysis software NVivo to openly encode, correlate and selectively encode the data, and finally tested the encoding results through the theoretical saturation test. **[Result/conclusion]** Finally, in the context of peer review of data papers, a scientific data quality evaluation index system was established, including four categories of data internal quality, data expression quality, data access quality, and data utility quality and 14 evaluation indicators. Then this article analyzed the specific meaning and classification of the indicators in detail to help the authors and reviewers of data papers understand the internal structure of scientific data quality.

Keywords: data journal scientific data quality peer review evaluation indicator grounded theory